(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 January 2002 (17.01.2002)

PCT

(10) International Publication Number
WO 02/05133 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/SG00/00100

(22) International Filing Date: 7 July 2000 (07.07.2000)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **KENT RIDGE DIGITAL LABS** [SG/SG]; 21 Heng Mui Keng Terrace, Singapore 119613 (SG).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LIM, Allison** [US/SG]; 360 Pasir Panjang Road #04-11, Singapore 118699 (SG). **WANG, Jiren** [CN/SG]; 4 Clementi Avenue, Block 315 #05-135, Singapore 120315 (SG). **WONG, Limsoon** [MY/MY]; 19 Jln I1, TMN Melawati, Kuala Lumpur 53100 (MY).

(74) Agents: **JACOB, Sheena, R.** et al.; Alban Tay Mahtani & de Silva, 39 Robinson Road, #07-01 Robinson Point, Singapore 068911 (SG).

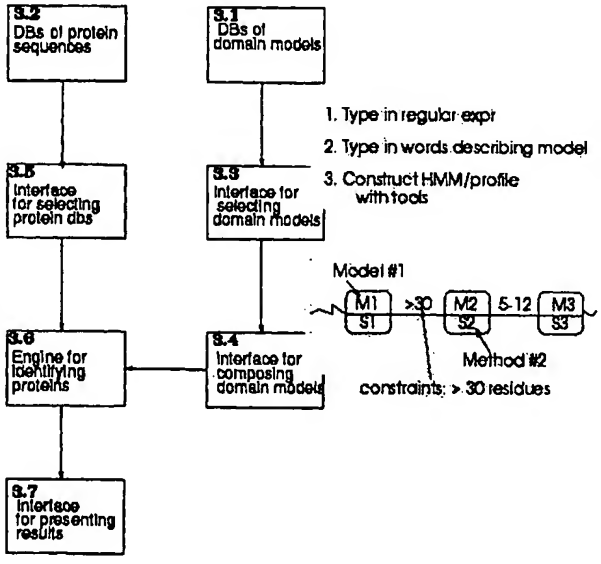
(81) Designated States *(national)*: SG, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: A METHOD AND APPARATUS FOR SEARCHING A DATABASE CONTAINING BIOLOGICAL INFORMATION



(57) Abstract: The present invention relates to a method and apparatus adapted to facilitate searching of a database containing biological information. The present invention, in particular but not exclusively, provides a search system and method that allows for flexible design of queries for sequences with certain biological function units (such as motif and domains) and identification of protein sequences that have these units. The design of queries, in one form, is based on a combination of existing models and/or user-defined queries to provide a flexible selection of criteria in defining a query or queries for interrogation of a database. The present invention has application, in one form, to the fields of bio-informatics, computer science, information science, pharmaceutical science and biotechnology.

WO 02/05133 A1

A METHOD AND APPARATUS FOR SEARCHING A DATABASE CONTAINING BIOLOGICAL INFORMATION.

FIELD OF INVENTION

The present invention relates to a method and apparatus adapted to facilitate searching of a database containing biological information. The present invention, in particular but not exclusively, provides a search system and method that allows for flexible design of queries for sequences with certain biological function units (such as motif and domains) and identification of protein sequences that have these units. The design of queries, in one form, is based on a combination of existing models and / or user-defined queries. The present invention has application, in one form, to the fields of bioinformatics, computer science, information science, pharmaceutical science and biotechnology.

BACKGROUND ART

Presently, in the state of the art, biological scientists express a need to identify proteins by their functional units, and which must satisfy certain compositional constraints.

Two examples of the current state of the art are:

1. to identify a special class of proteins that exhibit a special pattern within their zinc finger domains, and
2. to identify ``twinfilin'' proteins, which were proteins containing two copies of filin domain.

In both of these examples, existing off-the-shelf software products used for searching biological databases can not be used. The prior art software either does not support the domain models needed, does not support complex composition of domain models, or does not support both.

In the case of special zinc finger proteins, currently a scientist would need to run the hidden Markov model software HMMER [SR Eddy, ``Hidden Markov Models'', *Current Opinion in Structural Biology*, 6:361--365, 1996] on the zinc finger domain model from PFAM [E.L. Sonnhammer et al, "Pfam: A Comprehensive Database of Protein Families based on Seed Alignments", *Proteins*, 28:405--420, 1997] to pick out the preliminary zinc

finger proteins in a database. Then for each predicted zinc finger domain, Perl [E. Quigley, *Perl by Example*, Prentice Hall, 1994] would be used to test the predictions for the required pattern. In the case of twinfilin, the Entrez database [GD Schuler et al, ``Entrez: Molecular Biology Database and Retrieval System", *Methods in Enzymology*, 266:141—161, 1996] would be queried in order to extract examples of filin domain. A hidden Markov model of filin domain using these example sequences would then be constructed, and thereafter using this model, HMMER is used to identify proteins in the database that contain at least two non-overlapping significant hits.

Some prior art applications search data stored in a biological database representative of sequences for a user-defined amino acid or nucleotide pattern query. These applications (such as PatScan [R Overbeek, Argonne National Labs; http://www.mcs/anl/gov/compbio/PatScan], PattinProt [Pole Bio-Informatique Lyonnais; http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_pattern.html], ProSearch [LF Kolakowski et al, ``ProSearch: fast searching of protein sequences with regular expression patterns related to protein structure and function", *Biotechniques,* 13(6):919-21, 1992], PATTERN [Intelligenetics package, Oxford Molecular Ltd], QUEST [Intelligenetics package, Oxford Molecular Ltd], GeneMan [JP Clewley, ``GENEMAN of LASERGENE", *Methods Mol Biol.*, 70:189-96, 1997], Scrutineer [PR Sibbald et al, ``Scrutineer: a computer program that flexibly seeks and describes motifs and profiles in protein sequence databases", *Comput Appl Biosci.*, 6(3):279-88, 1990] , PatternFind [Bioinformatics Group, Swiss Institute for Experimental Cancer Research; http://www.isrec.isb-sib.ch/software/PATFND_form.html], FPAT [Institute for Biomedical Computing, Washington University at St. Louis; http://www.ibc.wustl.edu/fpat]) are limited to pattern searching using regular expression syntax.

A few other prior art applications (such as PATTERN MATCH [PIR International, http://www-nbrf.georgetown.edu/nbrf/scan.html], ScanProsite [Swiss Institute of Bioinformatics; http://www.expasy.ch/tools/scnpsite.html]) allow the user to identify a PROSITE motif (pattern) [Hofmann K et al, ``The

PROSITE database, its status in 1999", *Nucleic Acids Research*, 27(1):215—219, 1999], but this requires the user to know the accession number of the motif. Alternatively, the user can type in the motif whether self created or whether already defined in a motif library.

Still a few other applications (such as HMMER/PFAM [Bateman A et al, ``Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins", *Nucleic Acids Research*, 27(1):260—262, 1999], pfsearch [P Bucher et al, ``A flexible search technique based on generalized profiles", *Computers and Chemistry*, 20:3—24, 1996], PRINTS [TK Attwood et al, ``PRINTS prepares for the new millennium", *Nucleic Acids Research*, 27(1):220--225, 1999], EMOTIF [Nevill-Manning CG et al, ``Highly specific protein sequence motifs for genome analysis", *Proc Natl Acad Sci USA*, 95(11):5865-71, 1998], MAST [Bailey TL et al, ``Methods and statistics for combining motif match scores", *J Comput Biol.*, 5(2):211-21, 1998], BLOCKS [J. Henikoff et al, ``New features of the Blocks database servers", *Nucleic Acids Research*, 27(1):226—228, 1999]) use a hidden Markov model, position weight matrix profile, or blocks of alignment sequences to search for domains. In this case the user must explicitly provide a file of the hidden Markov model or profile. These applications also do not allow the user to specify composition of domains.

Another problem associated with the prior art is that many of the applications and tools do not let the user conveniently use them in combination with other domain identification applications.

Still further problems are considered to exist with the prior art. Existing methods are considered to be relatively inconvenient and inflexible, especially in as much as:

a. There is considered to be little help from the existing applications in formulating the query from known motif and domain libraries. A few applications allow the user to identify a pattern from PROSITE by the accession number only, but this is considered to be limiting. Otherwise, the user has to type in some kind of regular expression that is to be searched.

b. There is considered to be no variety in the type of query offered. There is also considered to be no readily accessible application for forming domain queries. In the prior art, usually, the applications only offer queries for patterns defined solely by regular expression syntax, solely by hidden Markov models or position weight matrix profiles (which must be created before hand). They are also considered not let different methods or tools to be combined and do not let multiple domain models to be composed.

c. Furthermore, there is considered to be a lack in flexibility in combining motifs/domains in query. The applications of the prior art are not considered to be able to flexibly combine motifs/domains from libraries with user-supplied patterns in one query to locate one or more motifs in the same sequence.

The present invention seeks as an object to alleviate at least one problem associated with the prior art.

SUMMARY OF INVENTION

The present invention, in one form, stems from the recognition that the problems a, b, and c noted above exist and a solution to the problems should be devised.

The present invention provides an interface/apparatus and /or method for devising a query for use in interrogating a biological database to identify a target protein, in which query a user is able to:

(a) describe a target protein's composition of domains,

(b) select at least one preferred means for identifying such domains,

(c) select at least one preferred protein database(s).

Preferably, the apparatus and / or method further serves to:

(d) execute the query by searching for the target protein by identifying those protein sequences from (c) having composition of domains from (a) detected using means from (b).

Advantageously, the present invention preferably includes a number of features, such as:

- English or human language description to select pre-defined domain models.
- Automatic construction of domain model from the human language description.
- Automatic derivation of domain model from user-supplied example sequences.
- Enabling multiple methods to be combined to search for the same domain (by imposing containment constraints on the domain models of these methods).
- Enabling multiple domains to be composed (by imposing distance constraints on these domain models).

The essence of the invention stems not so much from the constituent parts of the invention, but from the ability to combine and utilise a number of different systems or methods in order to obtain a flexible selection of criteria in defining a query or queries for interrogation of a database(s).

Many advantages are considered to arise from the present invention, including:

1. Assisting in formulating the query from a variety of pre-defined domain model databases, in which the present invention lets the user take advantage of the collection of domain model databases so that the user can search for a defined domain model by unique identifier or an English description.
2. An integration of a variety of search methods and tools. In the present invention, the user can utilize a variety of search methods associated with the domain model databases to execute the search.
3. The ability to combine pre-defined domain models with user-defined domain models in query. In the present invention, the user can combine domain models in pre-defined databases with his own domain models (which he supplies on-the-fly) of the same (or different) protein domain.
4. An ability to flexibly define multiple domain models (of different protein domains) in a query. With the present invention, the user can define complex compositions of domain models in one query. These types of queries can involve regular expressions, hidden Markov models, or position weight matrix profiles, and are particularly suitable for identifying complex multi-domain proteins. The query can also include other models or compositions as would be understood by those skilled in the art.
5. The ability to employ several search methods to achieve a more confident and reliable retrieval. This is achieved by searching for the same domain with different tools (i.e. different domain models and associated search methods) that have different coverage. For example, the user can specify that he wants to identify proteins containing a TPR domain [JR Lamb et al, ``Tetratrico peptide repeat interactions: to TPR or not to TPR?", *Trends in Biochemical Sciences*, 20(7):257—259, 1995] predicted by both hidden Markov model and

position weight matrix profile methods (by imposing a containment constraint on the two domain models).

6. In the present invention, the user can take advantage of the defined motif/domain libraries so that the user can search for a defined domain by a partial regular expression match or a keyword that describes the motif.

7. Also, in the present invention, the user can (a) use pre-defined domain models, (b) automatically create them using user-supplied English specification, and (c) automatically create them (both directly and indirectly) using user-supplied seed sequences. Moreover, the user can specify composition of models (eg. two filin domains) and combination of search methods (eg. occurrence of regular expression within a zinc finger domain).

Preferred Embodiment

A preferred embodiment of the present invention will now be described with reference to the accompanying drawings, in which:

Figure 1 illustrates schematically one embodiment of the present invention, and

Figure 2 illustrates an example output of the embodiment of Figure 1.

In one form, the present invention contains several major components described as follows:

a. An extensible collection of motifs, profiles, regular expression patterns, hidden Markov models, etc. with their associated search methods. These motifs, profiles, regular expressions, hidden Markov models, etc. are collectively referred to in this Disclosure as ``domain models". These domain models can be verbatim import from established external databases.

b. An extensible collection of databases of protein sequences. These databases may be existing library(s) or be compiled individually.

c. An interface allowing a user to select an individual domain model. The selection is achieved either by (a) entering an English description, and then selecting from matching entries in PROTEIN DESIGNER's

collection of domain models; or by (b) direct browsing of entries in PROTEIN DESIGNER's collection; or by (c) direct entry using regular expression; or by (d) direct derivation of hidden Markov model from a user-supplied list of seed protein sequences; or by (e) direct derivation of hidden Markov model from protein sequences in public databases matching a user-supplied list of seed protein sequences; or by (f) direct derivation of hidden Markov model from protein sequences in public databases matching a user-supplied English description.

d. An interface allowing the user to compose the individual models to form a description of the domain composition of the proteins he wishes to identify. This interface can be either graphical or text-based. The user uses it to specify (a) relative ordering of the domains in the target protein, (b) distance and/or containment constraints between these domains in the target protein, and (d) if necessary, scoring thresholds for these domains.

e. An interface allowing the user to select databases from PROTEIN DESIGNER's collection of databases.

f. An engine for applying the specified domain composition on the selected protein databases and for displaying the matching proteins.

In this embodiment, the present invention is referred to as the PROTEIN DESIGNER and provides a user a convenient way

(a) to describe a target protein's composition/arrangement of domains,

(b) to select preferred means for identifying such domains,

(c) to select preferred protein sequence databases, and

(d) to search for his target protein by identifying those protein sequences from (c) having composition/arrangement of domains from (a) detected using means from (b).

An embodiment of the PROTEIN DESIGNER is shown in the Figure 1, in which:

Embodiment of the ``DB of domain models" (3.1)

This is a list of domain models, associated thresholds, and English or other human language descriptions. The English or other human language descriptions are searchable. A standard relational database is preferably used in the implementation.

Embodiment of the ``DB of protein sequences" (3.2)

This is a list of protein sequences and their English descriptions. This list can be further divided into sublists based on the sources of these sequences or other criteria. A relational database or a FASTA-formatted flat file is again preferably used in the implementation.

Embodiment of the ``Interface for selecting domain models" (3.3)

In the interface 3.3, there are preferably a number of top-level options included, such as:

Option 1 is ``Type in regular expression". Under this option, the user is asked to provide a regular expression to specify the constituent model. The symbols allowed in this regular expression are the 20 amino acid letters and the dot symbol (.) representing don't-care. These symbols can be grouped using the square brackets ([and]) so that [ABC] means A or B or C. These symbols can be written adjacent to each other so that ABC means A followed by B followed by C. Each symbol can be annotated by a repetition constraint *{d}* meaning repeat *d* times; *{,d}* meaning repeat at most *d* times; *{d,}* meaning repeat at least *d* times; *{j,k}* meaning repeat between *j* to *k* times. For example, *C.{2,4}C.{3}[LIVMFYWC].{8}H.{3,5}H* specifies the usual zinc finger domain.

Option 2 is ``Type in English word". Under this option, the user is asked to provide a list of keywords describing a constituent domain. This keywords is then used to search the ``DB of domain models" (3.1) to find predefined domains whose description in the database contains these words. Those with more matching keywords are listed first. The user can then browse and select from this list a desired constituent domain model, provided that model appears on the list. A more sophisticated embodiment

can also make use of approximate matching of keywords based on stemming and thesaurus.

Option 3 is ``Construct HMM''. Under this option, there are several suboptions to let the user select the means for constructing a desired constituent domain model. The preferred, but not the only, means are the followings:

Direct derivation of hidden Markov model from a user-supplied list of seed protein sequences. This method works in the standard way as described in [R. Durbin, et. al. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, chapter 3, pages 46--79. Cambridge University Press, 1998.]: a multiple alignment of the seed protein sequences is computed; then the transition probabilities for each sequence position is computed; then the hidden Markov model is simply the log-likelihood of the sum of these transition probabilities as one moves from the initial sequence position to the last sequence position.

Direct derivation of hidden Markov model from protein sequences in public databases matching a user-supplied list of seed protein sequences. This method first uses each seed protein sequence to perform a BLASTP operation on a public database as described in [S. F. Altschul, et. al. ``Basic local alignment search tool'', *J. Molecular Biology*, 215:403--410, 1990.]; the aligned regions of the hits produced by each seed protein are combined into a single collection. The hidden Markov model is then derived as described earlier using this collection of protein (sub)sequences.

Direct derivation of hidden Markov model from protein sequences in public databases matching a user-supplied English description. The English words are used to search feature annotations in the public database Entrez [G. Schuler, et. al. ``Entrez: Molecular biology database and retrieval system'', *Methods in Enzymology*, 266:141--162, 1996.] to extract segments of protein sequences in Entrez corresponding to these matching annotations as described in [K. Lin, et. al. ``Hunting TPR domains using Kleisli'', *Genome Informatics Series*, 9:173--182, 1998. Universal Academy Press, Tokyo,

Japan.] The hidden Markov model is then derived as described earlier using this collection of protein (sub)sequences.

The user is also given an option to save the constructed hidden Markov model into the database (3.1). The options noted above may also be used partly or wholly in combination.

Embodiment of the ``Interface for composing domain models" (3.4)

Let us assume *M1*, ..., *Mn* are the constituent domain models selected from the ``Interface for selecting domain models" (3.3) and *S1*, ..., *Sn* are the corresponding methods/scoring thresholds associated with these selected constituent domain models. There are several conceivable ways for specifying how these domain models can be composed. We describe two ways below, merely for the purpose of illustration, and without limitation:

A textual method. We describe this method using a formal grammar. Let *SPEC* denote a domain composition specification and let *M* denote one of *M1*, ..., *Mn*. Then syntactically, a domain composition specification is formed by the following grammar:

		formation rule	Meaning
SPEC	::=	*M*	Any protein where *M* appears is acceptable
	|	*SPEC1* before{>d} *SPEC2*	Any protein where *SPEC1* appears before *SPEC2* and the occurrence of *SPEC1* and *SPEC2* is separated by a distance of at least *d* residues is acceptable
	|	*SPEC1* and{>d} *SPEC2*	Any protein where *SPEC1* and *SPEC2* both Appear and the occurrence of *SPEC1* and *SPEC2* overlaps by at least *d* residues is Acceptable
	|	*SPEC1* or *SPEC2*	Any protein where at least one of *SPEC1* or *SPEC2* appears is acceptable

Round brackets can be used to disambiguate where necessary.

A graphical method. An graphical icon is provided for each *Mi/Si* selected from ``Interface for selecting domain models" (3.3). A canvas is provided for the user to click and drop these icons. A line between two icons denotes ``before", in a left-to-right manner. A line can be annotated by a distance constraint *{>d}* and its means the constituent domain represented by the icon at its left is separated from the constituent domain represented by the icon at its right by at least *d* residues. A circle shaded in a light colour (say red) can be used to group icons. Such a circle means that all the constituents' domains represented by all the enclosed icons must appear in the desired protein. The circle can also be annotated by a distance constraint *{>d}* and its means that these constituent domains are expected to overlap by at least *d* residues. A circle shaded in a different light colour (say blue) can be used to group icons. Such a circle means that at least one of the constituent domains represented by the enclosed icons must appear in the desired protein. An example is shown in the shaded box adjacent to 3.4 in *Figure 1*.

The distance constraint *{>d}* above can also be generalized, for example, to *{j-k}* meaning at least *j* and at most *k*.

Embodiment of the ``Interface for selecting protein database" (3.5)

The names of the sublists from (3.2) are provided to the user for selection. Alternatively, the user can specify some keywords and all protein sequences whose English descriptions in (3.2) match these keywords are selected.

Embodiment of the ``Engine for identifying proteins" (3.6)

First, for each of the domain model and search method selected from the ``Interface for selecting domain models" (3.3), it applies the selected method on the selected model to the databases selected from the ``Interface for selecting proteins dbs" (3.5). A hit is defined as a domain predicted in a protein sequence at a score better than the threshold. The hits (the protein sequences, positions, and scores) found are saved. Second, it considers the distance constraints and containment constraints from the ``Interface for composing domain models" (3.4). A previously saved hit is eliminated of it fails any distance or containment constraint. A distance constraint between two domains fails if the positions predicted for these two domains does not satisfy the separation specified. A containment constraint between two domains fails if the positions predicted for these two domains does not satisfy the percentage overlap specified. Finally, all hits that succeed are saved for subsequent presentation by the ``Interface for presenting results" (3.7).

Embodiment of the ``Interface for presenting results" (3.7)

This interface looks up from the protein sequence database (3.2) the names, English or other language description, and sequence of each hit from (3.6). It then displays this information, together with a graphical or textual layout of the constituent domains of the corresponding hit.

Alternatively, this interface initially displays a summary of the hits. The summary contains just the name of the protein sequence in each hit, together with a graphical or textual layout of its constituent domains. The layout can be selected or defined or composed by the user .

An example presentation for results of a search consisting of three Tetratricopeptide repeat domains (TPR) is given in *Figure 2* . The

corresponding (textual) domain composition specification is "TPR before{0} TPR before{0} TPR". It shows three protein sequences (P14922, P30260, P38042) from the Swissprot database that satisfy the domain composition criterion.

5

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

1. A system for identifying protein sequences from large protein sequence databases by specifying the possible composition and arrangements of domains expected in them.

2. An apparatus for devising a query for use in interrogating a biological database to identify a target protein, the apparatus including:

representing means for describing the target protein's composition of domains,

first selecting means for selecting at least one preferred means for identifying such domains, and

second selecting means for selecting at least one preferred protein database(s).

3. An apparatus as claimed in claim 2, wherein the representing means enables English or human language description or a form-based description specifying the name(s) of the domain and (optionally) its expected length to be used to select pre-defined domain models.

4. An apparatus as claimed in claim 2, wherein the representing means enables automatic construction of domain model from an English or human language description.

5. An apparatus as claimed in claim 2, wherein the representing means enables automatic derivation of domain model from user-supplied example sequences.

6. An apparatus for executing a query as claimed in claim 2, 3, 4 or 5 on a database, the apparatus including execution means for executing the query by searching for the target protein by identifying those protein sequences from the second selecting means having composition of domains from the representing means detected using the first selecting means.

7. An apparatus as claimed in claim 5, including multiple search means for enabling multiple methods to be combined to search for the same domain and / or multiple domains to be composed.

8. An apparatus as claimed in any one of claims 6 or 7, further including graphically representing means for displaying the results of a query in either or both a graphical or text-based representation.

9. An apparatus as claimed in claim 8, wherein the display means can be configured to specify (a) relative ordering of the domains in the target protein, (b) distance and/or containment constraints between these domains in the target protein, and / or (c) if necessary, scoring thresholds for these domains.

10. An engine for applying a specified domain composition on selected protein database(s) and for displaying the matching proteins.

11. An engine as claimed in claim 10, including the apparatus of any one of claims 2 to 9.

12. An engine as claimed in claim 10 or 11, further including a database having an extensible collection of motifs, profiles, regular expression patterns, hidden Markov models, etc., collectively referred to as ``domain models", with their associated search methods and / or an extensible collection of databases of protein sequences.

13. An engine as claimed in claim 10, 11 or 12, further including means to enable a user to select individual domain model by:

(a) entering an English description, and then selecting from matching entries in PROTEIN DESIGNER's collection of domain models;

(b) direct browsing of entries in PROTEIN DESIGNER's collection;

(c) direct entry using regular expression;

(d) direct derivation of hidden Markov model from a user-supplied list of seed protein sequences;

(e) direct derivation of hidden Markov model from protein sequences in public databases matching a user-supplied list of seed protein sequences; and / or

(f) direct derivation of hidden Markov model from protein sequences in public databases matching a user-supplied English description.

14. An engine as claimed in any one of claims 10 to 13, further including means to enable a user to compose the individual models to form a description of the domain composition/arrangement of the proteins he wishes to identify by

relative ordering of the domains in the target protein,

(a) distance and/or containment constraints between these domains in the target protein, and

(b) if necessary, scoring thresholds for these domains, or

(c) a description involving disjunction of domain composition/arrangement.

15. An engine as claimed in any one of claims 10 to 14, further including:

(a) an additional database of precomputed domains obtained by applying all domain models to all protein databases, and / or

(b) each record of the additional database storing information on what domains are predicted by what methods at what positions at what scores in what protein sequences.

16. An engine as claimed in claim 15, in which the query is applied to the additional database of precomputed domains to locate proteins that satisfy a specified domain composition/arrangement.

17. A method for devising a query for use in interrogating a biological database to identify a target protein, the method including:

a. describing the target protein's composition of domains,

b. selecting at least one preferred means for identifying such domains, and

c. selecting at least one preferred protein database(s).

18. A method as claimed in claim 17, wherein English or a human language description or a form-based description specifying the name(s) of the domain and (optionally) its expected length is used to select pre-defined domain models.

19. A method of executing a query as devised according to claimed in claim 17 or 18, on a database, the method including the step of

executing the query by searching for the target protein by identifying those protein sequences from step c. having composition of domains from step a. detected using step b.

20. A method as claimed in claim 17, 18 or 19, further including allowing multiple methods to be combined to search for the same domain and / or multiple domains to be composed.

21. A method as claimed in any one of claims 17 to 20, further including the step of displaying the results either graphically or in a text-based format.

Figure 1

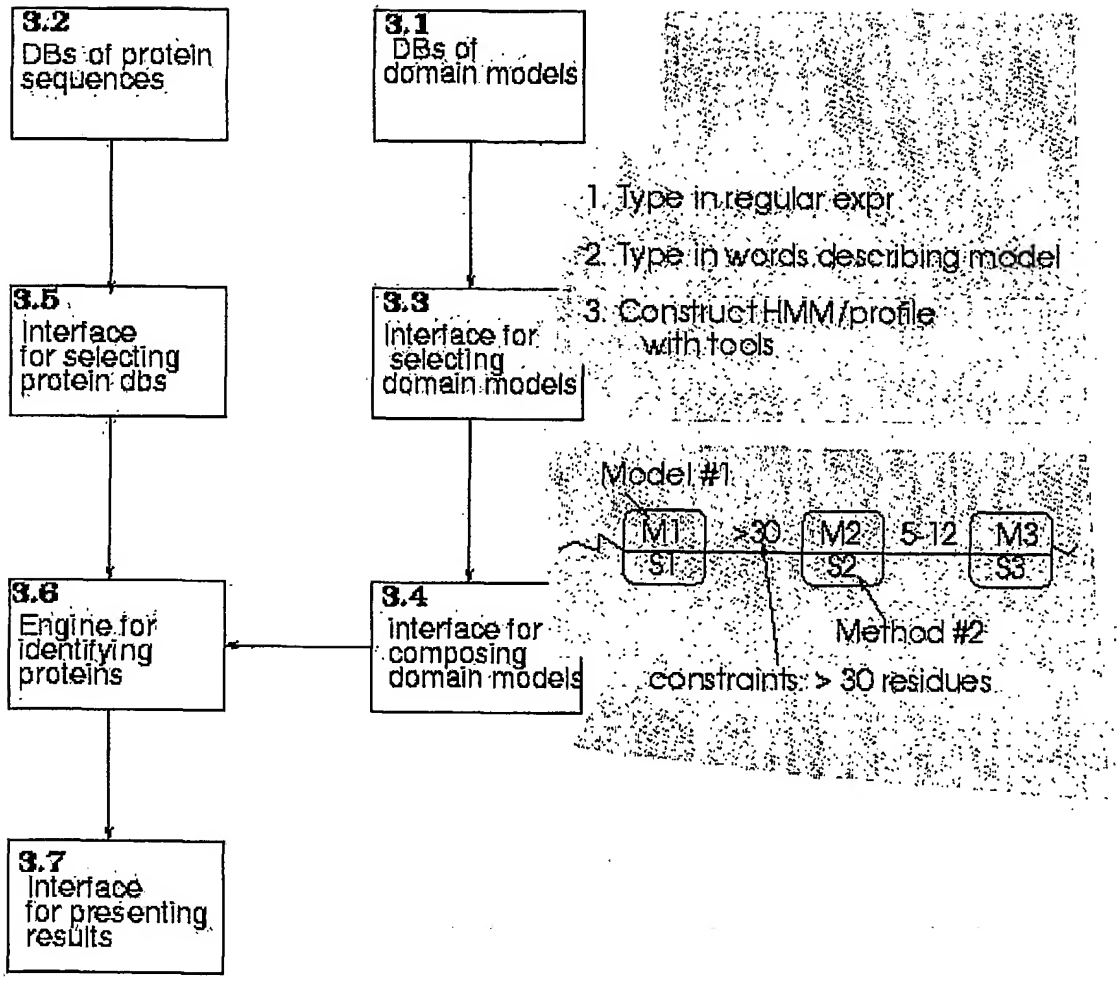
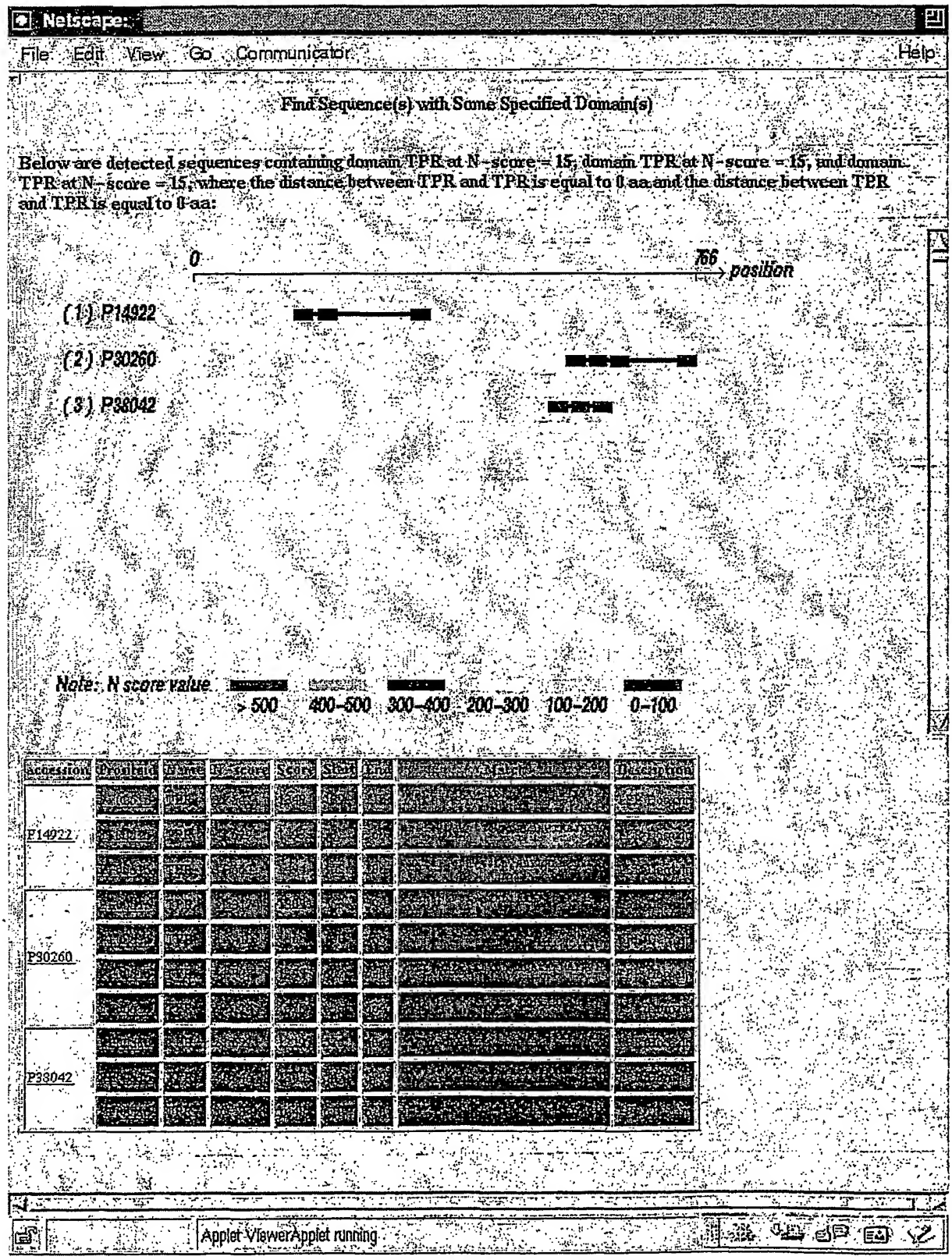


Figure 2



A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, COMPENDEX, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JASON WANG; BRUCE SHAPIRO; DENNIS SHASHA: "Pattern Discovery in Biomolecular Data" 1999 , OXFORD UNIVERSITY PRESS , NEW YORK OXFORD XP002168720 ISBN: 0-19-511940-1	1
Y	page 161, line 21 -page 162, line 8 page 165, line 12 -page 165, line 36 page 172, line 5 -page 173, line 21 page 175, line 1 -page 177, line 43 --- -/--	2-4,6, 8-13, 17-19,21

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
1 June 2001	19/06/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Boyadzhiev, Y

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CURWEN V A ET AL: "GHOST: a gene homology online search tool" TRENDS IN GENETICS,ELSEVIER, AMSTERDAM,NL, vol. 16, no. 7, July 2000 (2000-07), pages 321-323, XP004207252 ISSN: 0168-9525	2-4,6, 8-13, 17-19,21
A	the whole document	1
X	WO 00 23474 A (DOOLEY MICHAEL JOHN ;UNIV QUEENSLAND (AU); ANDREWS PETER RONALD (A) 27 April 2000 (2000-04-27)	1
A	abstract page 4, line 3 -page 4, line 24 page 7, line 27 -page 8, line 13 page 22, line 14 -page 22, line 25	2-21
A	WO 00 26818 A (RIGOUTSOS ISIDORE ;FLORATOS ARIS (US); IBM (US)) 11 May 2000 (2000-05-11) abstract page 5, line 1 -page 6, line 29 page 8, line 16 -page 11, line 24 page 14, line 9 -page 14, line 19 page 25, line 25 -page 26, line 9	1-21
A	ALTSCHUL S F ET AL: "Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases" TIBS TRENDS IN BIOCHEMICAL SCIENCES,EN,ELSEVIER PUBLICATION, CAMBRIDGE, vol. 23, no. 11, 1 November 1998 (1998-11-01), pages 444-447, XP004143492 ISSN: 0968-0004 the whole document	1-21

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0023474 A	27-04-2000	AU 1141100 A	08-05-2000
WO 0026818 A	11-05-2000	CN 1287641 T	14-03-2001
		CN 1289424 T	28-03-2001
		EP 1044417 A	18-10-2000
		EP 1057131 A	06-12-2000
		WO 0026819 A	11-05-2000